Introduction

This dataset reflects arrest incidents in the City of Los Angeles from 2010 to 2019. There are 1,320,000 rows in this dataset, each row represents an arrest. The questions are Whether the number of arrest incidents in the City of Los Angeles from 2010 to 2019 decreased, and is there any relationship of age with variables in sex, area, time and arrest type in the crime situation in 2019? The five main variables I choose represent criminal age, criminal sex, Patrol Divisions location, arrested time and the type of charge the individual was arrested for repectively.

Methods

In order to find out the answer, I have separated the dataset in 2010 and 2019. There are 162344 rows and 26 columns in crime data 2010, 88296 rows and 26 columns in crime data 2019. After check the main variables (Sex, Age, Area, Time and Arrest Type), I found there are missing values in the Time variable. So I imputed the missing value using the mean value grouped by sex, area and arrest type. I also found a case which the arrested position (LAT=0.00, LON=0.00) is too far away from the others, In order to make map looks more clear, I removed this case only in the mapping step. There is no implausible values in these main variables. I recoded the sex, arrest type and created a new variable "Part of Day" based on Time for the analysis in the next step. After data wrangling step, I created the summary tables for mean of age, standard deviation of age, female propotion and male propotion grouped by area, arrested type and part of day both in 2010 and 2019.

For the difference between 2010 and 2019, I choose to make barcahrt to show the variation in arrest type, criminal sex, part of day, and boxplot to show the variation in criminal age. I also made a map to show difference in the distribution of arrested position. However the map for the whole dataset in these two years are too slow to show up (the dataset is too large), I took the distribution of arrested position arrested by Patrol Divisions in central area as an instance.

For the relationship of age with sex, time, area and arrest type in 2019, I used the histogram to show the correlations between age with arrest type and part of day. I used summary graphs to show the correlations between age with sex and area. Finally I used anova function to calculate the p-value and showt the relationship statistically.

Preliminary Results

From the first step, I found that the total number of arrest cases are dramatically droped. In 2010, there are 129346 males are arrested and 32998 females. While in 2019, there are only 69491 males and 18805 females are arrested, almost half. In 2010, the mean and median age of criminals are 32.16 and 29 while in 2019 increase to 35.25 and 33

respectively. Misdemeanor is the most common type of arrest. There are 106249 criminals in 2010 were arrested because of that while only 48533 in 2019. Patrol divisions in Hollywood has the most arrest cases, in 2010 (15671), while in 2019, Central area became the most. Arrest incidences decrease from 0:00 to 5:00, arrive its lowest point at about 5:00 and highest point at about 16:00 both in 2010 and 2019.

Results table

	Difference with 2010
Number of cases	Decreased
Arrest Time	Almost no change
Criminal Sex	Almost no change
Patrol Divisions Area	arrested position are more concentrated near the patrol division area
Criminal Age	Older
Arrest Type	Relatively more proportion of Felony

Summary of Results

Variables	F-value	p-value	Related with age?
Arrest Type	783.1	<0.05	Yes
Arrest Time	467.2	<0.05	Yes
Criminal Sex	857.6	<0.05	Yes
Patrol Divisions Area	146.5	<0.05	Yes

brief Conclusion
According to my questions in the first step, the arrest incidences in the City of LA in 2010 and 2019 are different in the total number, criminal age, patrol divisions area and arrest type. In 2019, the age of arrest criminal in the City of LA is correlated in arrest type, arrest time, criminal sex, and patrol divisions area.